第 37 卷第 3 期 计算机应用研究 Vol. 37 No. 3
录用定稿 Application Research of Computers Accepted Paper

基于条件随机场与信息熵的特定领域概念发现 *

付　瑶 [1]，万　静 [1†]，邢立栋 [2]

(1. 北京化工大学 信息科学与技术学院，北京 100029; 2. 中国科学院自动化研究所，北京 100190)

摘　要：针对特定领域内自动化识别既有概念和发现新概念的问题，提出了一种基于条件随机场和信息熵的抽取方法。通过使用条件随机场对文本中的概念词进行边界预测，与词典中的概念对比，筛选出新概念的候选项并找出其大概位置，然后由互信息和左右熵分别判断概念窗口内的概念内部结合度和概念边界自由度，从而发现新的专业概念。实验表明，使用该方法进行概念发现比单独使用条件随机场的方法有更好的效果，基于字和词的模型概念发现的准确率分别提升了 20.06%和 46.54%。

关键词：概念识别；新概念发现；条件随机场；信息熵；特定领域

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.08.0623

Crf and information entropy based method for new words discovery in specific domain

Fu Yao[1], Wan Jing[1†], Xing Lidong[2]

(1. *College of Information Science & Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*; 2. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Aiming at the problem of automatic identification of existing concepts and discovering new concepts in a specific field, a method based on conditional random field and information entropy is proposed. The conditional random field is used to predict the boundary of conceptual words in text. The candidates of the new concept can be selected with the comparison to the existing concepts in the dictionary and the probably location in text is found. Then the mutual information and the left and right entropy are used to judge the internal degree of integration and the boundary freedom of the concept in the concept window for discovering new professional concepts. Experiments show that the concept discovery using this method has a better effect than the method of using the conditional random field alone. The accuracy of the concept discovery based on word and words model is respectively improved by 20.06% and 46.54%.

Key words: concept recognition; new concept discovery; conditional random field; information entropy; specific field

0　引言

近几年，科学研究的快速发展使得各领域的专业词汇层出不穷。通常情况下，特定领域的专业词汇大多常出现在对应领域的知识传播媒介中，它们有较多的特殊性和专业性，只有专业人员才会对这些词汇有一定的了解，而新的领域概念出现的速度随着研究的进步甚至已经超过了部分领域学者的认知速度。因此，如何高效、准确、全面地识别与发现专业领域的新词，具有非常重要的意义。

在相关研究工作中，专业领域新词发现的方法主要分为基于规则的方法和基于统计的方法[1]。基于规则的新词发现方法需要构建规则库，即领域专家根据专业知识的发展以及语言学原理制定各领域构词的共性和个性规则，并依此进行新词发现。李明[2]利用改进后的 Apriori 算法对语料处理并生成关联规则，然后利用生成的规则对新的专业词汇进行抽取工作。Sasano 等人[3]针对日语中出现的新词，利用衍生规则和象声词模式，通过在句子的格式框架中添加新节点的方式发现最优路径，以此实现对新词的识别，该方法对某些特定类别的新词有很好的识别效果。郑家恒等人[4]根据汉语构词法建立规则库，通过调用“互斥性字串”过滤规则和构词规则发现新词。基于规则的方法发现新词的准确率较高，但受领域限制严重，专业词汇的更新速度快，需要不断地更新既有规则，构建成本较高。基于统计的新词发现方法是通过大量的语料计算词频、词共现概率等统计学特征来识别领域新词。Li 等人[5]提出了一种基于词内部结合度和边界自由度的方法，对分词产生的“散串”处理进行新词发现。夭荣朋等人[6]使用 N-Gram 算法得到候选新词，再通过改进互信息和邻接熵对候选项扩展和过滤，结合词典筛选得到新词。Lei 等人[7]提出了一种层次聚类方法，将微博语料划分成具有不同主题的组，加强新词的统计特征，从而提高对新词提取的准确性。基于统计的方法不受领域限制，但由于数据稀疏，通常情况下识别的准确率不高。

针对两种方法各自的缺点，不少学者提出了基于统计和规则的方法。杜丽萍等人[8]利用互信息的改进算法与少量基本规则结合，实现了从语料中自动地识别网络新词，通过基于百度贴吧语料的实验，说明了该方法在大规模语料中发现新词的有效性。雷一鸣等人[9]采用互信息统计模型加入向右邻元迭代的方法进行新词候选集的获取，并通过引入外部统计量的概念对低频词进行过滤筛选得到新词。周霜霜[10]采用人工启发式规则对微博新词进行分类和归纳，再通过使用改进的 C/NC-value 算法融合 CRF 和 SVM 模型，提高了新词边界识别的准确率和低频新词识别的精度。基于统计和规则的方法结合了两者的优点，在新词识别方面往往可以取得比较好的效果。

收稿日期：2018-08-17；**修回日期：**2018-10-19　**基金项目：**国家科技支撑计划资助项目（2015BAK03B04）

作者简介：付瑶（1995-），女，黑龙江齐齐哈尔人，硕士研究生，主要研究方向为知识图谱、推荐算法；万静（1975-），女（通信作者），副教授，博士，主要研究方向为知识图谱、信息抽取（wanj@mail.buct.edu.cn）；邢立栋（1992-），男，工程师，硕士，主要研究方向为自然语言处理、知识图谱.

目前，多种机器学习方法也被应用到新词发现的任务中，并取得了较好的识别效果，如 CRF（条件随机场）、HMM（隐马尔可夫模型）、SVM（支持向量机）、DT（决策树）等。陈飞等人[11]归纳了许多区分新词边界的统计特征，利用 CRF 方法并综合这些特征在 SogouT 大规模语料上进行新词发现实验。丁祥武等人[12]在医疗领域文本结构化的过程中，使用 word2vec 将文本中的词转换为向量，通过词向量之间的得分高低表示词内部结合度的大小，再结合左右熵、词频等统计信息发现医疗文本中的新词。徐远方等人[13]将候选新词与词特征向量化得到新词候选向量，并与训练得到的支持向量构建矩阵，通过 SVM 测试得到新词。

本文提出一种新的基于 CRF 模型结合互信息与左右熵的特定领域新词发现方法。首先由既有的专业词汇对语料进行标注，训练条件随机场；然后由训练得到的模型来识别出候选字符串，这时候选字符串中部分是完整的专业词汇，部分是不完整的专业词汇，其余为非专业词汇，过滤掉词表中已有的概念，剩余的字符串通过互信息对这些词进行拼接，通过左右熵对这些词进行筛选，得到新的概念。

1　相关理论

1.1　条件随机场

条件随机场（conditional random fields，CRF）[14]是一种判别式概率模型，是马尔可夫随机场的一种。它可以使用复杂、有重叠性和非独立的特征进行训练和推理，既能充分利用上下文信息作为特征，也可以添加其他的外部特征。通过该模型训练能够获取丰富的特征信息，同时可以解决最大熵模型中的标注偏置等问题。

CRF 模型如图 1 所示。顶点间的连线代表随机变量间的相依关系。在条件随机场中，随机变量 Y 满足条件概率分布，给定的观察序列为随机变量 X 。

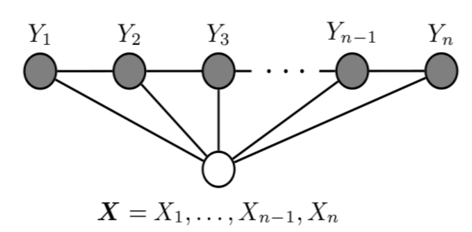


图 1　CRF 图解

Fig. 1　Graphic of CRF

设观察序列 $X=\{X_1,X_2,X_3\cdots X_n\}$ ，这里的输入数据可以是文本中的字或者词等。其对应的状态序列为 $Y=\{Y_1,Y_2,Y_3,\cdots,Y_n\}$ 。当给定观察序列 X 的取值 x 时，状态序列 Y 取值为 y 的条件概率的计算方法如下：

$$p(y\mid x)=\frac{1}{Z(x)}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \tag{1}$$

其中： θ_k 是对应特征函数 f_k 的权重参数；特征函数中的 y_t 、 y_{t-1} 分别表示文本当前输出状态和上一个输出状态； x_t 是当前输入状态； $Z(x)$ 是归一化因子。其计算方法为

$$Z(x)=\sum_{y}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \tag{2}$$

在应用 CRF 的过程中，特征的选择直接影响到特征函数的有效性。特征的选择没有固定的形式，要根据目标领域、文本语言、文本表述特征等方面进行综合考虑。通常情况下会将输入状态序列特征叠加组合。

1.2　互信息

互信息（mutual information，MI）[15]通常用来衡量两个随机变量之间的相互依赖程度。互信息越大，说明该二元组成为新词或者新词的一部分的可能性越大，即词内部结合度越大，通过与预设的阈值比较，当词内部结合度大于阈值时，即认为两者可以构成词语。两个随机变量 x 和 y 的互信息定义为

$$MI(x,y)=\log_2\frac{p(xy)}{p(x)p(y)} \tag{3}$$

其中： $p(xy)$ 是 x 和 y 的联合概率分布函数，即两者在语料里同时出现的概率； $p(x)$ 和 $p(y)$ 分别是 x 和 y 的边缘概率分布函数，即各自单独在语料中出现的概率。当 $MI(x,y)>>0$ 时，表明 x 和 y 是高度相关的，即 x 和 y 经常同时出现，字符串 xy 构成新词的可能性更大；当 $MI(x,y)=0$ 时，表明 x 和 y 是相互独立分布的；当 $MI(x,y)<<0$ 时，表明 x 和 y 是互不相关的。

1.3　左右熵

左右熵即词语的左右邻接熵（branch entropy，BE）[16]，可以用来衡量词语左右邻接字符的不确定性。语料中一个有意义的词语往往会有较高的频率出现在不同文档中，具有较高的灵活性，即可以与各种不同的外部条件进行搭配，搭配的种类越多，说明这个词语越灵活，边界自由度越高。本文引入候选新词的左右熵作为新词边界自由度的量化手段。候选词 w 的左右熵分别定义为

$$H_l=-\sum_{w_l\in s_l}p(w_l\mid w)\log_2 p(w_l\mid w) \tag{4}$$

$$H_r=-\sum_{w_r\in s_r}p(w_r\mid w)\log_2 p(w_r\mid w) \tag{5}$$

其中： s_l 是候选词 w 的左邻接字集合； w_l 是 s_l 中的元素； s_r 是 w 的右邻接字集合； w_r 是 s_r 中的元素。如果候选词的左右熵都较大，则说明与该候选词左右相邻的词串种类较多，相邻词频率分布较均匀，候选词与相邻词构成新词的概率较低；如果候选词的左右熵中有一个较小，则表示与该候选词相邻的不同词串的频率分布并不均匀，此时，候选词与相邻频率较高的词串组成新词的概率较高。

2　基于条件随机场与信息熵的特定领域概念发现

本文将特定领域的概念发现视为预测语料文本序列边界的问题。将概念发现融合于语料分词的过程中，对比既有词表发现新概念。方法主要由语料标注、CRF 模型训练和候选概念识别、互信息拼接与左右熵筛选三部分组成，如图 2 所示。

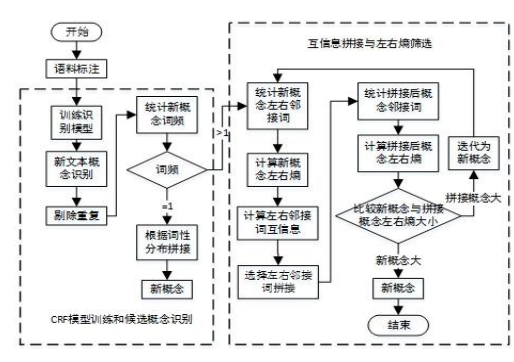


图 2　算法流程

Fig. 2　Algorithm flow

2.1　语料标注

本文所用到的标注集如表 1 所示。所采用的是分词常用的 BEMSN 标注集，确定词首、词中、词尾、单个词与无关词。此外，分词需要标注的特征是词性，本文的词性标注详情参照《HanLP 词性标注集》。

表 1　标注集

Table 1　Annotation set

标注描述	标注符号
概念词首	B
概念词中	M
概念词尾	E
单个概念词	S
无关词	N

2.2　CRF 模型训练和候选概念识别

数据处理过程如图 3 所示，概念识别可以看做序列化数据的标注问题。利用分词工具对原始语料进行分词，并充分利用上下文信息和其他外部特征，分词的结果按照词表中的概念标注词性、词首、词中和词尾，输入为带有词性标记的词序列，如式（6）所示。

$$WT = w_1/t_1\ w_2/t_2 \cdots w_n/t_n \tag{6}$$

其中：n 表示句子被切分得到的词的个数；t_i 表示词；w_i 表示被标注的词性。

利用标注的语料对 CRF 模型的参数进行训练，通过基于 CRF 模型的学习过程，得到领域概念的识别模型。

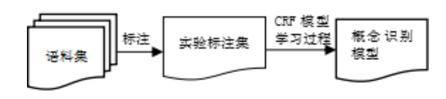


图 3　训练模型图

Fig. 3　Training model graph

通过得到的概念识别模型，结合 CRF 解码算法，对新的语料文本进行概念词边界的识别工程，即对部分词语进行拆分组合，最后输出一个最优的“词形/词性”序列 WC^*/TC^*，用式子可表示为

$$WC^*/TC^* = wc_1/tc_1\ wc_2/tc_2 \cdots wc_i/tc_i \cdots wc_n/tc_n \tag{7}$$

其中：$i \le n$，$wc_i = [w_j \cdots w_{j+k}]$，$tc_i = [t_j \cdots t_{j+k}]$，$1 \le k, j+k \le n$。该过程如图 4 所示。

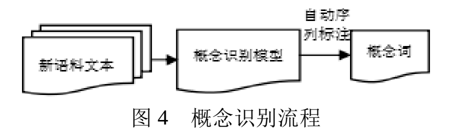


图 4　概念识别流程

Fig. 4　Extraction of concept graph

2.3　互信息拼接与左右熵筛选

经过条件随机场模型的识别得到候选概念词，其中不正确的新概念词多是由于字符串缺失不完整，本文提出的算法将对这些不正确的概念进行编辑。其基本思想是：统计不正确的概念词的左右候选词，计算相应的互信息，选取互信息值较大者拼接得到新词汇，再计算该新词汇的左右信息熵，取左右熵中的较小值作为新词的信息熵。如此循环递归，可得约束条件下，信息熵取最大值的新词即为发现的新概念。例如，通过条件随机场识别出候选概念词——“施工过程仿真”，这是一个不完整的词，通过互信息拼接右词后得到新词——“施工过程仿真分析”，计算“施工过程仿真分析”的左右信息熵，取左右熵中的最小值为其信息熵，与拼接过程中出现的其他候选词的信息熵进行比较，发现“施工过程仿真分析”的信息熵最大，即“施工过程仿真分析”为发现的新概念。具体算法如下：

输入：候选概念词集合 $S = \{S_1, S_2, S_3, \ldots, S_n\}$。

输出：信息熵最大时的概念词 $S_i^j = \arg\max(H_j)$。

1. foreach $i = \{1,2,3 \ldots n\}$ **do** 　// K_i 为 S_i 的词频

2. 　**if** $K_i = 1$ 　// $pos(S_i^j)$ 为词性分布拼接方法

3. 　　**return** $S_i^j = pos(S_i^j)$；

4. 　**else if** $K_i > 1$ 　// w_l，w_r 分别为 S_i^j 左右邻接词，$mi(w)$ 为求互信息方法

5. 　　**foreach** $j = \{1,2,3 \ldots n\}$ **do**

6. 　　　**if** $mi(w_l) > mi(w_r)$

7. 　　　　$S_i^j \leftarrow w_l\ S_i^{j-1}$；

8. 　　　else

9. 　　　　$S_i^j \leftarrow S_i^{j-1}\ w_r$； 　// $set(s)$ 为存储 S_i^j 方法

10. 　　　set(S_i^j);

11. 　　**end for** 　// $\arg h(s)$ 为求 $set(S_i^j)$ 中元素左右熵最大值时 S_i^j 的值

12. 　　**return** $S_i^j = \arg h(set(S_i^j))$；

13. end for

3　实验设置及结果分析

3.1　实验数据集

本文以建筑工程领域的图书期刊为语料集，对文中提出的方法在该语料集上进行了新词发现实验。共提取了 70 962 个建工领域概念词汇，并利用这些概念标注了 245 本建筑工程领域图书。

图 5 为原始数据的样例。图书中的所有文本作为本文实验语料。

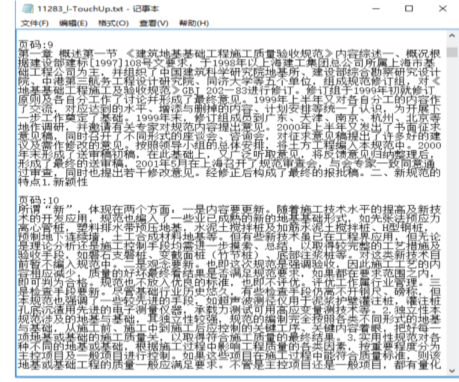


图 5　建筑工程图书文本

Fig. 5　Book text of construction project

为了从字符和词汇两个不同的分词粒度研究信息抽取，本文将实验语料分为两组：一组利用 HanLP[1]分词工具对初步抽取的结果进行分词，并依照《HanLP 词性标注集》附加词性信息，另一组则直接以字符作为实验数据。

本文采用如图 6 中的格式整理实验数据，其中每一种标注的第一列是待识别词，第二列是词性特征，第三列是正确标注。

3.2　实验方案

本文使用特征模板辅助生成特征函数，模板文件如图 7 所示，其中的每一行是一个模板。每个模板都由%x[Row, Column]来指定输入数据中的每一个单元。Row 代表当前单元的行偏移，Column 代表列位置。

实验采用 CRF++0.58[2]作为 CRF 的实现工具。为了方便表述，将前面提到的特征集分别用字母做标志，如表 2 所示。

[1] HanLP 是开源的汉语言处理包 http://hanlp.linrunsoft.com/index.html

[2] CRF++是著名的条件随机场开源工具 http://crfpp.sourceforge.net

```
本书　r N
较　d N
全面　ad N
较　d N
系统　n S
地　ude2 N
叙述　v N
了　ule N
湿陷性　nz B
黄土　n E
地区　n N
工业　n N
与　cc N
民用　b B
建筑　n E
的　n N
设计　vn N
与　cc N
施工　vn N
，　w N
是　vshi N
我国　n N
西北　ns N
地区　n N
三十　m N
多年来　d N
地基　n B
与　cc M
基础　n M
工程　n E
的　n N
经验　n N
总结　v N
。　w N
```

图 6　基于词的建工概念标注

Fig. 6　Word based construction project's concept labeling

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

U23:%x[0,1]

# Bigram
B
```

图 7　特征模板示意图

Fig. 7　Characteristic template

表 2　特征表示方法

Table 2　Feature representation method

词性特征	POS
互信息拼接	MI
左右熵筛选	EN
以词为基本单位	WORD
以字为基本单位	CHAR

本文共做了四组交叉实验对上文所提算法的有效性进行验证。

实验 1：　基于字或词的 CRF 对比实验。研究 CRF 模型结合不同粒度的标注方式（词或者字）对原始文本做信息抽取工作的效果；

实验 2：　词性特征有效性判定实验。在以词为基本单元的实验组中，通过加入词性特征判断其对抽取效果的影响；

实验 3：　加入互信息与左右熵交叉对比实验。在以字或词为基本单元的实验组中，采用加入互信息和左右熵的方法进行实验，以对比其与未加入互信息和左右熵时对概念识别的影响；

实验 4：融合实验。利用实验二中的识别模型标注新文本识别新概念，利用互信息与左右熵进行拼接与筛选，通过与其他实验组做对比，判断词性特征与识别后加入互信息、左右熵对文本识别的影响。

3.3　实验结果与分析

实验结果如图 8 所示。图中 *WP* 表示新词发现的识别准确率，*WR* 表示新词发现识别召回率，*WF* 分别表示新词发现的 *F* 值，*NP* 表示建筑工程领域概念的识别准确率，*NR* 表示建筑工程领域概念的识别召回率，*NF* 表示建筑工程领域概念识别的 *F* 值。

基于字或词的 CRF 对比实验结果如图 8 所示。即实验 1 表明，以字作为识别的基本单元时，准确率和召回率都比以词作为基本单元时高。采用 CRF 模型，并以字作为基本单位时，其召回率在新词发现或概念识别上分别比以词作为基本单位的最好识别水平高 3.56%或 5.7%。

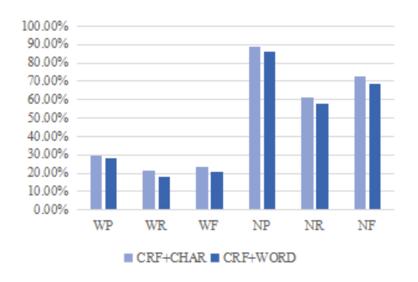


图 8　字与词对比实验

Fig. 8　Character and Word contrast experiment

词性特征有效性判定实验结果如图 9 所示。实验 2 表明，以词作为基本单元，加上词性特征比未加入词性特征，在信息抽取的召回率上有提高。

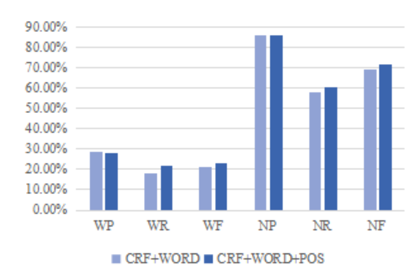


图 9　加入词性特征的影响

Fig. 9　Effect of joining part of speech feature

实验 3 的结果如图 10 所示。数据表明，在识别后加入互信息和左右熵，概念发现的效果有明显提升。基于字的模型在识别后加入互信息和左右熵，其准确率提升了 20.06%，基于词的模型在识别后加入互信息与左右熵，其准确率提升了 46.54%。经对比发现，前者在识别效果的提升上低于后者，分析认为是基于字的模型拼接字不如基于词的模型直接拼接词得到的结果更完整，从而影响了效果。

融合实验，即使用本文提出的概念识别方法进行的实验，通过加入了词性特征的条件随机场模型，提升了识别的准确率和效率，减少了运行所需要的时间，找到了新概念的大概位置，加入词性分布并利用互信息和左右熵进行拼接筛选，取得了很好的效果。实验结果如图 11 所示。

小结：利用互信息与左右熵进行条件随机场识别后处理可以有效提高概念发现工作的准确率及召回率。条件随机场识别的作用在于发现概念所在的大概位置，基于这个位置利用信息论的方法可以提取出完整准确的建工领域概念。此外，基于字的识别方法与基于词的识别方法对比，通过条件随机场模型发现的概念位置基本一样，经过信息论方法处理后其

识别的准确率和召回率前者低于后者，另外，加入词性特征的基于词的模型在识别后加入互信息与左右熵，处理得到的结果效果最好。

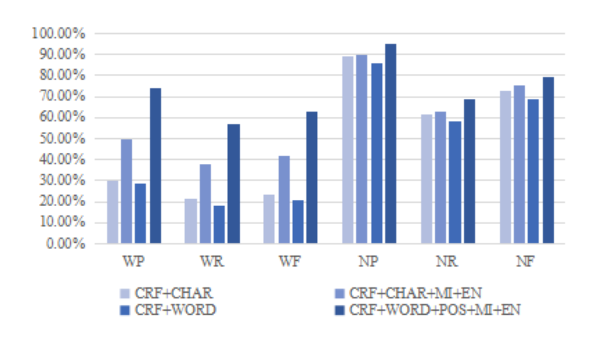


图 10　加入互信息与信息熵影响

Fig. 10　Effect of joining mutual information and information entropy

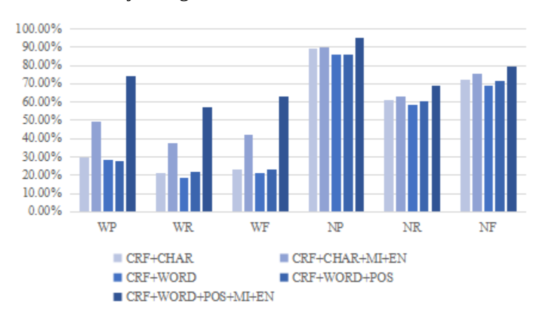


图 11　多特征融合实验

Fig. 11　Multi feature fusion experiment

4　结束语

本文提出了基于条件随机场与左右熵的特定领域概念识别方法，即对语料进行标注，由处理过的数据训练条件随机场模型，并用其识别候选词，通过词表过滤掉已有概念，对其余候选词使用互信息和左右熵进行拼接筛选，得到新的概念。实验使用建工领域的语料集，并分别从分词粒度、特征选择、加入互信息和左右熵处理等维度验证了文本提出的方法在建筑工程领域的概念发现与识别中的有效性。实验结果表明，该方法在不增加人工标注的条件下提高了信息抽取的准确率和召回率，同时可以提高识别的效率。另需注意，标注集的质量、机器的性能等会很大程度上影响识别模型的训练，从而影响概念识别的效果。

参考文献：

[1] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述 [J]. 计算机科学, 2010, 37 (3): 6-10. (Zhang Haijun, Shi Shumin, Zhu Chaoyong, *et al.* Survey of Chinese new words identification [J]. Computer Science, 2010, 37 (3): 6-10.)

[2] 李明. 针对特定领域的中文新词发现技术研究 [D]. 南京：南京航空航天大学, 2012. (Li Ming. New words discovery research for specific areas [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2012.)

[3] Sasano R, Kurohashi S, Okumura M. A simple approach to unknown word processing in Japanese morphological analysis [C]//Proc of International Joint Conference on Natural Language Processing. 2013: 162-170.

[4] 郑家恒，李文花. 基于构词法的网络新词自动识别初探 [J]. 山西大学学报：自然科学版 , 2002, 25 (2): 115-119. (Zheng Jiaheng, Li Wenhua. A study on automatic identification for internet new words according to word-building rule [J]. Journal of Shanxi University:Na t. Sci. Ed. , 2002, 25 (2): 115-119.)

[5] Li Wenkun, Zhang Yangsen, Chen Ruoyu. New word detection based on inner combination degree and boundary freedom degree of word [J]. Application Research of Computers, 2015, 32 (8) , 2302-2304.

[6] 夭荣朋，许国艳，宋健. 基于改进互信息和邻接熵的微博新词发现方法 [J]. 计算机应用, 2016, 36 (10): 2772-2776. (Yao Rongpeng, Xu Guoyan, Song Jian. Micro-blog new word discovery method based on improved mutual information and branch entropy [J]. Journal of Computer Applications, 2016, 36 (10): 2772-2776.)

[7] Lei K, Zhang W Y, Zhang K, *et al.* Extracting unknown words from Sina Weibo via data clustering [C]//Proc of IEEE International Conference on Communciations. 2016: 1182-1187.

[8] 杜丽萍，李晓戈，于根，等. 基于互信息改进算法的新词发现对中文分词系统改进 [J]. 北京大学学报：自然科学版 , 2016, 52 (1): 35-40. (Du Liping, Li Xiaoge, Yu Gen, et al. New word detection based on an improved pmi algorithm for enhancing segmentation system [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52 (1): 35-40.)

[9] 雷一鸣，刘勇，霍华. 面向网络语言基于微博语料的新词发现方法 [J]. 计算机工程与设计, 2017, 38 (3): 789-794. (Lei Yiming, Liu Yong, Huo Hua. New word discovery based on microblog corpus for network language [J]. Computer Engineering and Design, 2017, 38 (3): 789-794.)

[10] 周霜霜. 基于规则与统计相融合的微博新词发现研究 [D]. 北京：北京交通大学, 2017. (Zhou Shuangshuang. Combinating of rules and statistics for new word detection of microblog text. Beijing: Beijing Jiaotong University, 2017.)

[11] 陈飞，刘奕群，魏超,等. 基于条件随机场方法的开放领域新词发现 [J]. 软件学报, 2013, 24 (5): 1051-1060. (Chen Fei, Liu Yiqun, Wei Chao, et al. Open domain new word detection using condition random field method [J]. Journal of Software, 2013, 24 (5): 1051-1060.

[12] 丁祥武，张夕华. 医疗领域文本结构化 [J]. 计算机工程与设计, 2017, 38 (10): 2873-2878. (Ding Xiangwu, Zhang Xihua. Text structuralization in medical field [J]. Computer Engineering and Design, 2017, 38 (10): 2873-2878.)

[13] 徐远方，李成城. 基于 SVM 和词间特征的新词识别研究 [J]. 计算机技术与发展, 2012, 22 (5): 134-136. (Xu Yuanfang, Li Chengcheng. Research on new word identification based on svm and word characteristics [J]. Computer Technology and Development, 2012, 22 (5): 134-136)

[14] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of International Conference on Machine Learning, Massachusetts. 2001: 282-289.

[15] Ong T H, Chen H . Updateable PAT-tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management [C]// Proc of the 2nd Asian Digital Library Conference. 1999: 63-84.

[16] 刘伟童，刘培玉，刘文锋,等. 基于互信息和邻接熵的新词发现算法 [J/OL]. 计算机应用研究, 2019, 36 (5) . (Liu Weitong, Liu Peiyu, Liu Wenfeng, *et al.* New word discovery algorithm based on mutual information and branch entropy [J/OL]. Application Research of Computers, 2019, 36 (5) .) (http://www.arocmag.com/article/02-2019-05-017. Html.)